

Weekly Report (2018.9.17-2018.9.23)

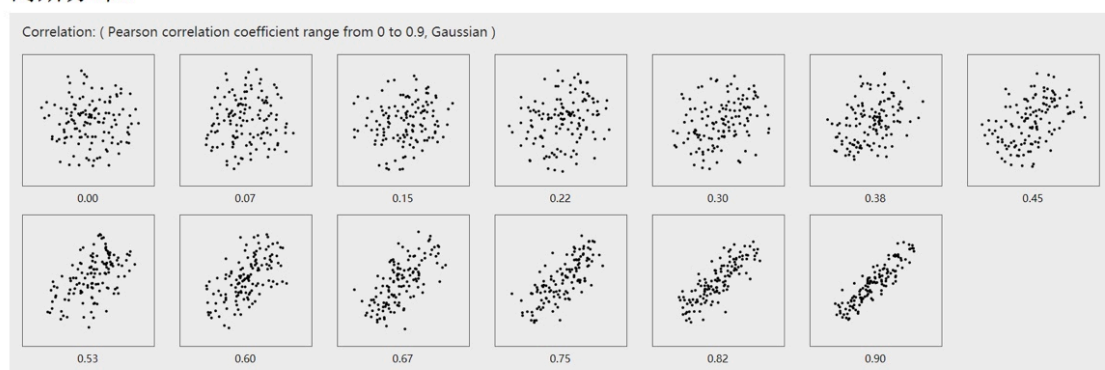
DONE

1. CHI2019投稿:

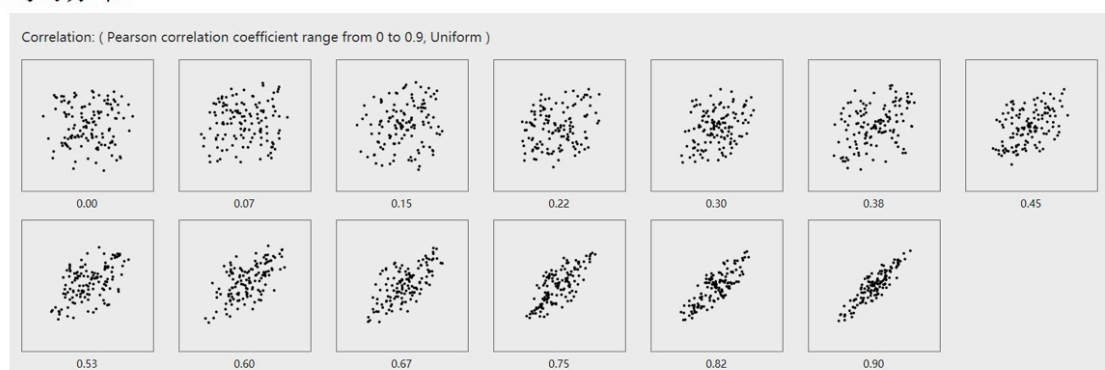
- 实验实现与用户测试: 周初进行了两次 pilot study, 根据实验结果对实验的参数做了相应调整, 然后自己在做实验的过程中发现画出来的correlation与cluster的散点图质量不高, 对数据生成方案进行了修改。在实验反馈中, 我们自己以及其他同学会发现一些系统上的问题, 包括交互、自动化等方面的问题, 也一并进行了修改。在修改过这些问题后, 进行了中等规模的 pilot study, 得到12组结果数据, 接下来要做结果分析。
- 数据生成方案修改: 对于 correlation, 通过看绘制出来的相关性的图, 发现点很偏, 对程序进行修改, 发现参考的其中一篇文章的公式打错了, 少了一个平方, 这是通过另一篇文章里的同一个公式发现的。修改过后发现画出的点整体还是会往上偏一些, 公式是正确的, 这就没办法了, 于是就放弃使用公式, 直接使用高斯分布的二元联合分布去做, cov里面的rho刚好是皮尔逊相关系数, 兜兜转转回到最初的方案。至于均匀分布, 尝试了几个想法, 最后决定在椭圆里面均匀撒点, 来尽量模拟。对于 cluster separation, 抛弃之前的直接使用中心点间距离来生成散点图数据, 那个做法显得有些草率, 最开始只是简单做了一个程序, 看到轮廓系数与中心点间距离大致呈现线性关系, 就直接选择了使用相对简单的中心点间距离来生成数据, 有些不够严谨。借鉴相关系数数据的生成方法, 可以先使用一系列的点间距离 (不局限于之前设置的那13个值) 生成数据集, 然后计算其轮廓系数, 满足某一个轮廓系数就留下, 否则舍弃, 继续循环, 这里在判断是否留下时, 需要设置一个可接受的误差范围来提高留用的概率, 至于误差范围的确定, 因为轮廓系数的取值范围[0,0.65]与相关系数的取值范围[0,0.9]相近, 所以误差范围的设置同相关系数, 为0.005 (来自几篇做相关系数的工作)。
- 目前生成的数据的效果图:

Correlation

高斯分布

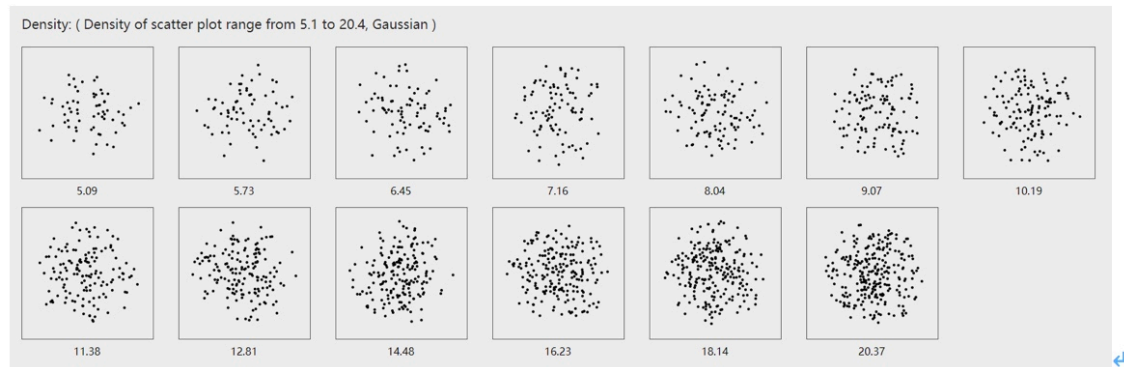


均匀分布

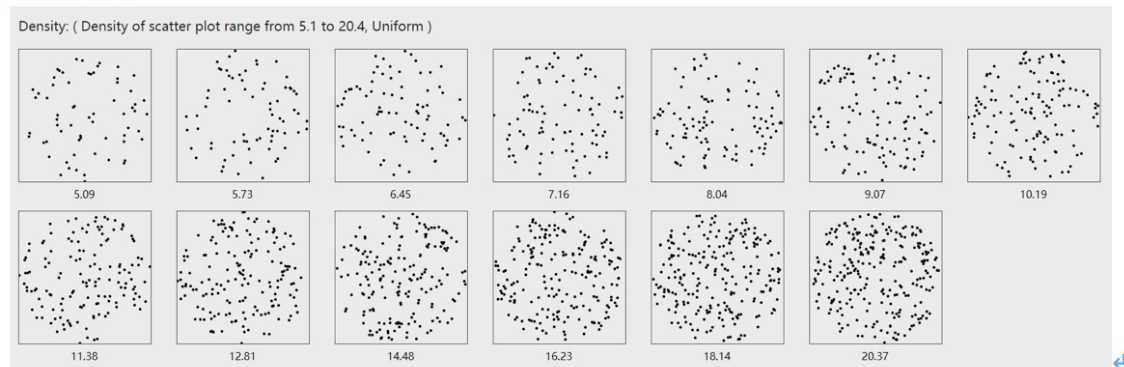


Density

高斯分布

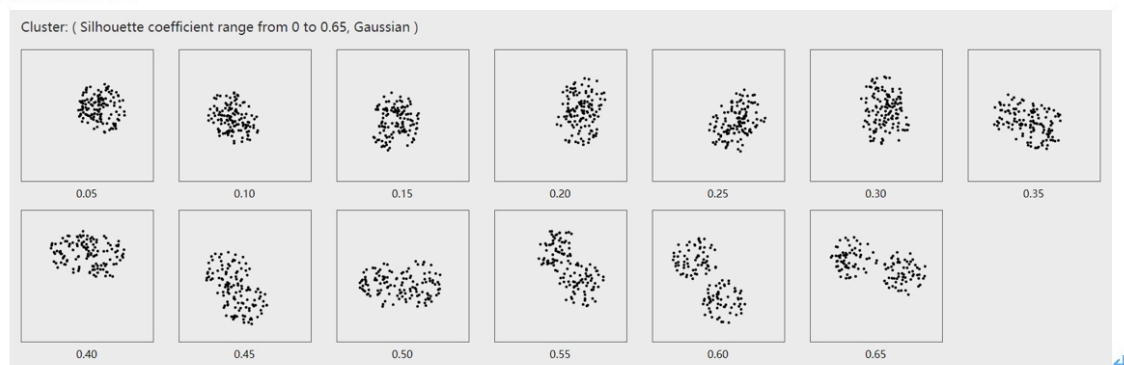


均匀分布

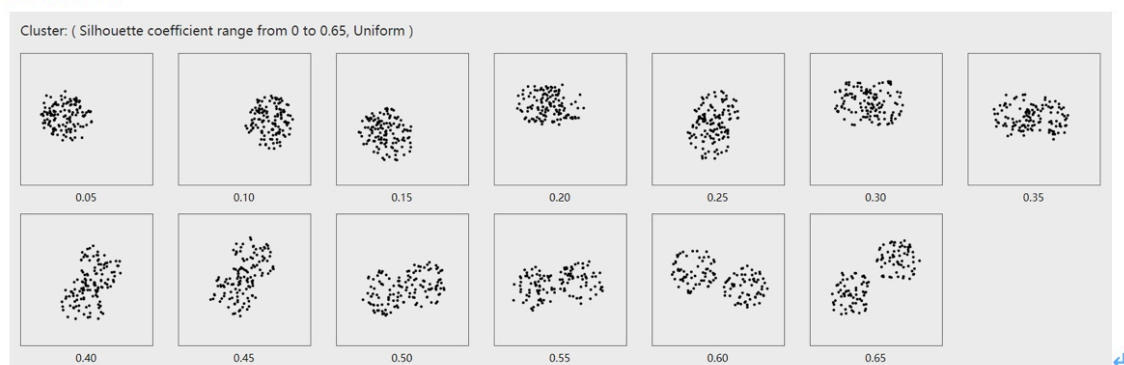


Cluster

高斯分布



均匀分布



2. RSATree

- 代码交接：只是简单的与元哲交接了一下程序后端，没有开始这个系统的编码。

小结

工作日工作时长9.5-10.5h，周末4小时，总时长约54h。这周修改数据生成代码的时间比较多，每一个参数几乎都改了几版的方案，直到最后的较好的效果，在这过程中，锻炼了自己的 python 编程能力，学到很多。这周由于膝盖摔伤，没有运动，每日工作时间都有所加长，下周要加入锻炼时间。

PLAN

短期计划

1. 对得到的预实验结果进行分析，根据结果也许还会对实验进行修改，实验所有参数都验证无误后开始正式试验，届时这个项目可以告一段落。

中期计划

1. RSATree 投稿项目：重新整理项目整体代码，补充测试部分的代码，后面需要重新做一系列对比实验。
2. Visevo 论文。
3. 动态图查询调研。

长期计划

1. 了解更多机器学习、数据挖掘相关的算法。
2. 在项目中锻炼自己的思考能力与代码能力。